The Challenge of Analyzing Conversational Speech

Johannes Luig (johannes.luig@student.tugraz.at)

Seminar Work for "Advanced Signal Processing II", June 2008

Abstract

This paper gives an overview of the four main challenges arising when analyzing conversational (spontaneous) speech and presents recent approaches to cope with *Hidden Punctuation*, *Disfluencies*, *Turn-Taking* and *Emotional Speech*.

1 Hidden Punctuation

1.1 Written vs. Spoken Language

Written language provides explicit punctuation – word boundaries are marked by white spaces, sentence boundaries by commas and full stops. In contrast, spoken language is a continuous stream of words without any obvious lexical marking; the phrasing is mainly expressed through prosody (i.e., the "rhythm" of the word stream).

1.2 Importance of word and sentence boundaries

These word and sentence boundaries are, however, of great importance for automatic downstream processing tasks like parsing, information extraction, dialog act modeling, summarization and translation as well as for human readability of text-to-speech application outputs.

Furthermore, the performance of the speech recognition algorithm itself is increased by distinct segment boundaries.

A "common" language model is created by recording someone reading written text. Assuming proper articulation, this process implies an acoustical segmentation into sentence-like units, which can be implemented by simply chopping the speech signal at longer pauses and speaker changes.

Conversational speech introduces two problems here: on the one hand, a pause is neither a necessary nor a sufficient indicator of a sentence boundaries. On the other hand, several sentences are strung together without pauses.

1.3 Endpointing

Endpointing is a particular task, where the aim is to determine when an utterance is "complete". It is applied in online human-computer dialog systems and therefore a real-time procedure, i.e., it must be accomplished using information only before the potential endpoint.

The common method to solve this problem is to simply wait for a pause that is longer than pre-defined threshold – which is obviously not a very successful approach when dealing with arbitrarily distributed pauses.

1.4 A Prosody-based Approach

1.4.1 Concept

An approach introduced by [5] does not consider a single pause threshold, but additionally defines a set of "decision points" (depicted in Fig. 1), which cover shorter time intervals and serve as trigger points for prosodic analysis.

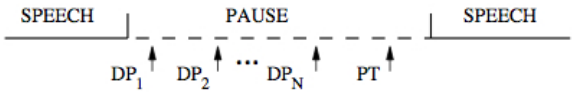

Figure 1: Multiple threshold concept. (from [5])

If a "pause" is detected (which is, in this case, defined as a silent period longer than $30ms$) and its duration exceeds the first decision point (DP_1), prosodic feature extraction is performed, and a score is created from this model for the first DP. This cycle is repeated, until the score exceeds a certain score threshold or if the pause threshold is reached in the meantime.

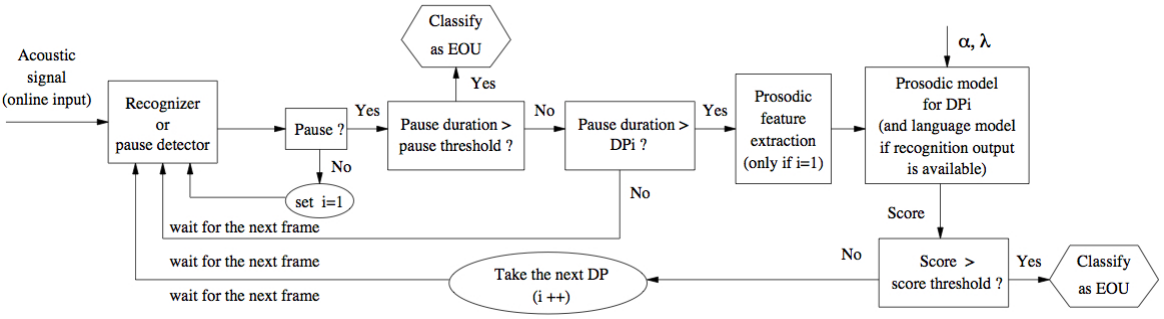

Figure 2: Flow Diagram of End-of-Utterance Detection (from [5])

1.4.2 Features

The above-mentioned approach uses duration- and pitch-based features. The duration features are extracted from time alignments delivered by the recognizer output; a typical example would be the duration of the last *rhyme*, i.e., the distance between the beginning of the last vowel and the end of the word.

The pitch features are speaker-specific pitch parameters as, for example, the speaker's "floor" (which is the lowest f_0 value he/she reaches); in this case, the distance from the average pitch in the last word to the speaker's "floor" could serve as a feature derived from the fundamental frequency.

2 Disfluencies

2.1 Types of Disfluencies

When talking about "disfluencies", one can distinguish the following different types:

Repetitions – The speaker repeats some part of the utterance: *I ... I like it.*

Revisions – The speaker modifies some part of the utterance: *We ... I like it.*

Restarts – The speaker abandons an utterance and then starts over: *It's also ... I like it.*

Filled Pauses – The speaker comments his editing: *We ... I mean, I like it.*

Across many languages, disfluencies occur at rates higher than every 20 words and can affect up to one third of all utterances.

2.2 Structure of Disfluencies

Disfluencies can be broken down into three regions: the *reparandum*, an optional *editing phase* and the *resumption*. In the examples above, the dots denote the right edge of the reparandum region, which is called the *interruption point* (IP).

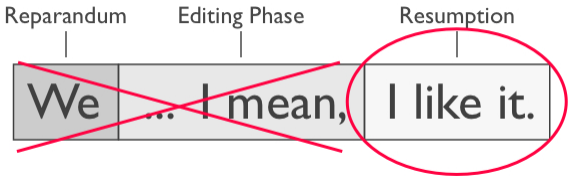

Figure 3: The three regions of a disfluency.

For speech recognition purposes, we solely need the resumption part, which means that we have to determine the interruption point (as well as the extent of the "cutaway" region for filled pauses). The latter is definitely a non-trivial task, since some words can function either as a "filler" or as a "non-filler".

2.3 An Approach based on Prosodic Features and Language Models

The following method has been proposed by [2]. It tries to Þrst detect the interruption point, and then apply some knowledge-based rules to identify the disfluency starting point (i.e., the begin of the reparandum region); filled pauses are not considered in this case. The corresponding system diagram is shown in Fig. 4.

In order to detect interruption points, so-called *Hidden-Event Language Models* are used. In such a model, each event is represented by an additional non-word token, for example: *I <IP> I like it.* The event token <IP> is explicitly represented and included in the vocabulary of the n-gram LM.

Three different models are incorporated into the system: a word-based LM models $P(W, E)$, the joint distribution of the event sequence E and the word string W; a second model based

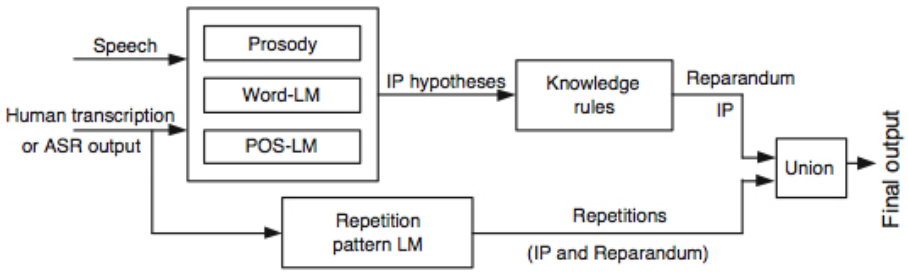

Figure 4: Flow diagram of Disfluency Identification. (from [2])

on "part-of-speech" (POS) tries to capture syntactically generalized patterns, such as the tendency to repeat prepositions that have occurred frequently in the training data.

In parallel, a special "Repetition Pattern LM" looks for repeating word strings explicitly.

3 Turn-Taking

Turn-taking is an expression for the speaker alternation in conversations, which – unfortunately – does not happen sequentially as in a play script. In fact, listeners anticipate the end of a speaker's contribution by analyzing syntax (word order), semantics (meaning of words), pragmatics (more than explicitly said) and prosody (word rhythm), and start talking before current speaker is finished.

3.1 Approach: Spectrogram Decomposition

A proposal by [6] is to decompose the spectrogram of a single-channel recording featuring several speakers by modeling each STFT magnitude vector as the outcome of a discrete random process that generates frequency bin indices.

The distribution of this random process is modeled as a mixture multinomial distribution; the specific multinomial distribution for each speaker is learned from training data (supervised approach). Source separation is obtained by computing maximum likelihood estimates of the speaker probabilities and the mixture weights.

4 Emotions

Emotion recognition and modeling is an inherently difficult task for a machine, since it requires more than just words. Rhetorical devices like irony (where "what is said" is the opposite of "what is meant") require interpretation and contextual relation.

Beyond, "natural" emotion data is very rare, as emotional databases are mostly created from acted speech; and the labeling of gathered data is difficult even for humans.

Emotion recognition has to deal with two intractable problems:

- Finding the classification reference, i.e., the "neutral" state
- Setting the "emotion analysis time window", since human emotions change over time

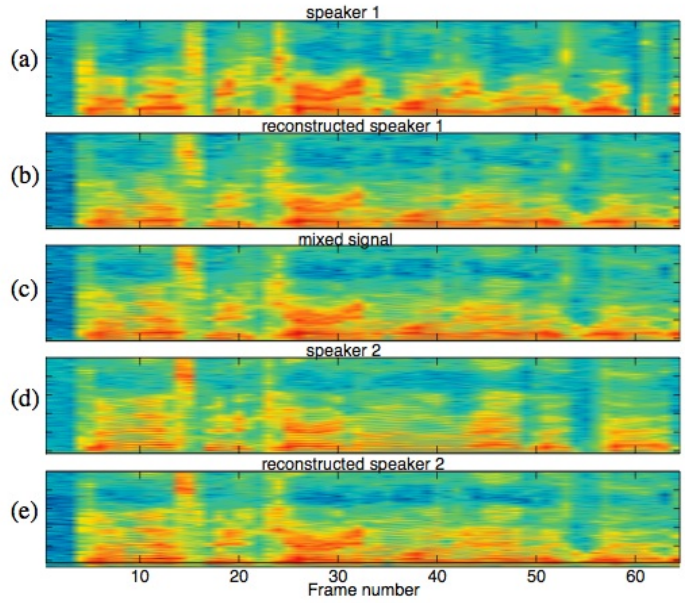

Figure 5: Spectrograms of original (a,d) and reconstructed (b,e) signals for two speakers and mixed signal (c). (from [6])

4.1 Features used for Emotion Analysis

"Classic" features used in this field include duration and rhythmic aspects such as speaking and pause rates (time domain) as well as pitch- and energy-based features (frequency domain).

Recent Approaches introduced, among others, the following features to determine emotional content:

- Loudness (RMS) across critical bands (Bark scale)
- Glottal-excitation-derived features (e.g. the "glottal volume velocity" signal, which is obtained by pitch-synchronous inverse-filtering of each glottal cycle of the waveform)
- Voice quality features (Jitter, Shimmer, etc.)
- Harmonicity features

5 Conclusion / Outlook

In this paper, we have become acquainted with the four main challenges concerning the analysis of conversational speech. Approaches to cope with the diverse problems mainly incorporate prosodic features and extended language models.

Results have not been discussed, as they are hardly comparable due to completely different topics; however, concerning accuracy and efficiency, the full potential of the discussed methods has for sure not been tapped yet.

Improved basic features, the incorporation of "longer-range" information (greater time windows) and speaker-dependent modeling not only in frame-level acoustics, but also in lexical and prosodic patterns will be great improvement opportunities.

References

[1] Elizabeth Shriberg: "Spontaneous Speech: How People Really Talk and Why Engineers Should Care", in *Proc. EUROSPEECH, Lisbon (Portugal), 2005*

[2] Yang Liu, Elizabeth Shriberg, Andreas Stolcke: "Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources", in *Proc. EUROSPEECH, Geneva (Switzerland), 2003*

[3] Anton Batliner et al.: "How to Find Trouble in Communication", in *Speech Communication, 40, 2003*

[4] Nick Campbell: "Conversational Speech Synthesis and the Need for Some Laughter", in *Journal of LaTeX Class files, Vol. 1, November 2002*

[5] Luciana Ferrer, Elizabeth Shriberg, Andreas Stolcke: "A Prosody-based Approach to End of-Utterance Detection that does not require Speech Recognition", in *Proc. ICASSP, 2003*

[6] Bhiksha Raj, Paris Smaragdis: "Latent Variable Decomposition of Spectrograms for Single-Channel Speaker Separation", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2005*